

EMC breidt productaanbod uit met analytische database for Big Data

Greenplum, polymorfe data-opslag, cloud en meer

Hans Lamboo

San Mateo, California, heeft zich in de IT-wereld nadrukkelijk op de kaart gezet. Het blijkt de afgelopen jaren een broedplaats te zijn voor vele nieuwe vormen van databases. Ook Greenplum, een analytische database die de 'polymorphic storage' introduceert, heeft er zijn roots. EMC lijfde het bedrijf nog niet zo lang geleden in.

Uwe Weimar, VP Central and Eastern Europe bij EMC/Greenplum, verdeelt de databasemarkt in twee hoofdtypen: OLTP en de analytische. Greenplum focust geheel op dat laatste. "Ons bedrijf begon met 70 man, en telt er nu bijna 1000 wereldwijd. Twee jaar geleden hadden we 100 klanten, maar dat waren dan ook hele goede klanten," vertelt hij. "In datawarehousing zijn de beslissingstrajecten, de lead times, lang. Ik schat tussen de 9 maanden in de VS en de 36 maanden in Duitsland. Dat vergt omzichtig manoeuvreren en vooral geduld." Greenplum focust op vier marktsegmenten; degene met de hoogste potentie is volgens Weimar de Overheid. "Maar daar zijn de beslissingstrajecten zo nog veel langer. We doen graag zaken in de Telecom, omdat Telco's hightech bedrijven zijn en meestal snel kunnen beslissen, in de retail, en in de financiële sector. Kort geleden ontdekten we een nieuwe markt voor Greenplum: de bedrijven die hun geld verdienen op internet. Die hebben gigantische hoeveelheden data, die ze moeten analyseren om aan de eisen van hun klanten te kunnen en blijven voldoen."

Niet rustig slapen

Wat betreft de technologische prestaties van Greenplum heeft Weimar, de salesman die hij nu eenmaal is, een expliciete mening: "We hebben de snelste analytische database ter wereld. Daar ben ik absoluut van overtuigd. De reden dat ik dat zonder blikken of blozen kan zeggen, ligt in het feit dat Greenplum maximaal schaalbaar is. Er is geen enkele andere leverancier die dat kan zeggen. We gebruiken onze unieke *shared nothing massive parallel* technologie. De combinatie van de schaalbaarheid en MPP levert een database op die supersonisch snel is." Een partij als Teradata claimt hetzelfde, maar ook daar heeft Weimar een antwoord op. "Ze hebben misschien de snelste *appliance*. Hoewel ik dat ook betwijfel, maar dat terzijde," zegt hij. "Wij hebben de snelste *database*. Ons systeem is volledig

gebaseerd op open source software, op PostgreSQL. Dat draait op elk type hardware, onder Linux, Unix, het maakt niet uit. Dat alles maakt de gebruiker volkomen onafhankelijk. Teradata komt als *appliance*, als een prachtige NCR-doos met een ingebouwde database. En daar zit direct het nadeel: je kunt de database niet scheiden van de software. Net als bij Oracle's Exadata. Je zit er dus met je ziel en zaligheid aan vast."

Greenplum is een zelfstandige business unit binnen EMC, een vergelijkbare positie als VMware. De filosofie van EMC is dat ze snelle groei die Greenplum doormaakt niet willen frustreren. "Die groei was gestopt toen Greenplum nog een zelfstandig bedrijf was. Hoe dat kwam? Als ik bij potentiële klanten kwam, vaak hele grote bedrijven, liepen de gesprekken eigenlijk allemaal op dezelfde manier af. Ze waren ervan overtuigd dat Greenplum de snelste database ter wereld is, met een betaalbare prijs. Maar wie is Greenplum? Bestaat dat bedrijf volgende week nog wel? Moeten we in zee gaan met een clubje van 70 man? Teradata levert ons een half zo goede performance tegen het dubbele van de prijs, maar we weten zeker dat ze morgen nog bestaan. Daarna kon ik mijn spullen pakken en met stille trom vertrekken. Nu we onder de vleugels van EMC opereren is dat probleem als sneeuw voor de zon verdwenen. EMC staat bekend als een goede, serieuze en stabiele onderneming, die al decennia lang in de markt actief is. We zien nu dat die combinatie ons elk jaar meer naar de rechterbovenhoek van Gartner's Magic Quadrant drijft." Om daaraan toe te voegen: "Als ik concurrent was van Greenplum, zou ik niet rustig slapen."

Consolidatie

Bart Sjerps, Advisory Technology Consultant – EMEA bij EMC, ziet dat Greenplum voor EMC een strategische functie heeft. "Greenplum is geen *point solution*, niet iets dat alleen maar leuk is om te hebben. Nee, het is heel strategisch in de markt van Big Data. EMC beweegt zich in twee gebieden. De een is die van *rich media*, de ander Analytics. En volgens IDC zit de grootste groei in internet-content met op de tweede plaats Analytics." De markt consolideert al enige tijd. EMC groeide over de laatste jaren van 40.000 naar 56.000 mensen. "EMC heeft in het verleden bewezen erg serieus om te gaan met investeringen. Kijk maar naar RSA, naar Documentum, naar VMware. Zeer succesvolle bedrijven die min of meer zelfstandig



Uwe Weimar: "Denk wat de combinatie van Hadoop en Greenplum allemaal kan".

opereren onder de paraplu van EMC," stelt Weimar. Hij vertelt: "Onlangs ontmoette ik een van de directeurs van Campfire, een van de divisies van vlaggenschip VMware. Campfire vormt het in-memory gedeelte van een database. In mijn ogen is het gewoon een plug-in. Het houdt de belangrijkste data vast in RAM, die daardoor extreem snel benaderd kunnen worden, bijna real-time. In het geheugen staat alleen het belangrijkste, de rest staat in het datawarehouse. Het bepalen wat de meest belangrijke data zijn, dat moet de consultant doen. Maar Campfire heeft alle benodigde filters om het mogelijk te maken." De door Greenplum toegepaste Massive Parallel Processing is geen exclusiviteit van dit bedrijf. Weimar: "Dat is zo. Maar Greenplum is gebaseerd op PostgreSQL. Open source, dat zo van het web kan worden gedownload. MPP voor iedereen, dus. Al 30.000 keer werd onze Greenplum Community Edition inmiddels al gedownload. Misschien op zich geen extreem hoog aantal, maar wel voor zo'n specialistisch product als dit. Met de Community Edition kun je alles doen, zonder beperkingen. Gratis. Je kunt bouwen en ontwikkelen, je kunt er zelfs

BusinessObjects, Cognos, MicroStrategy en SAS bovenop laten draaien. De enige limitering is dat je maar twee sockets kunt gebruiken."

Sjerps wil nog iets toevoegen aan het verhaal over MPP. "Kijk eens naar de huidige Intel-architectuur en naar SMP-machines. Als je nu één niveau verder gaat, bijvoorbeeld naar geclusterde databases, dan is de vraag: wat is de snelste transactie? Volgens mij is dat de transactie die je het dichtst bij de CPU doet. Als ik nu een machine bouw met meerdere CPU's en daar een database op laat draaien – dan ga ik er vanuit dat multi-threading en dat soort zaken zijn ingebouwd – dan moeten de CPU's met elkaar communiceren. 'Ik doe nu dit', 'ik moet nu daar een lock doen' enzovoort. Datzelfde geldt voor een niveau hoger, namelijk tussen servers.

Hier in Nederland zijn goede markten voor retail, olie en gas en telco's

Wat gebeurt er nu als je dat allemaal terzijde schuift en gewoon elke CPU een klein gedeelte van de database laat draaien? Dus ik heb een machine met 64 CPU's en laat elke CPU 1/64 van de database draaien. Die databases weten niets van elkaar, maar dat hoeft ook helemaal niet. Ze doen alles op hun eigen manier met hun eigen data. Dat gaat supersnel, want er is geen SMP- of clustering-overhead, niets. Dat noem ik échte MPP. Dan is schaalbaarheid een peulenschil. En dat is een van de fundamentele principes van Greenplum's architectuur. Vergeet niet: het gaat om standaard-CPU's. En dat is het verschil met bijvoorbeeld Teradata en Netezza."

Sjerps realiseert zich dat dat vooral komt omdat de CPU's een jaar of wat geleden niet krachtig genoeg waren, zodat de ontwikkelaars van appliances naar andere manieren moesten zoeken om de performance op te krikken. "Tegenwoordig verdubbelen Intel's CPU's elk twee jaar. Greenplum hoeft helemaal niets te doen. Onze database wordt automatisch sneller als de CPU's sneller worden. De leveranciers van appliances daarentegen moeten elke twee jaar nieuwe ASIC's ontwikkelen." Kortom, Greenplum is alleen software, die gebruik maakt van standaard hardware en standaard CPU's. Er is één master server die alle segment-servers aanstuurt. "Als de gebruiker 16 servers heeft, dan zijn die over twee jaar oud en moeten er nieuwe gekocht worden. Technisch gezien is er geen enkele beperking om de nieuwste generatie Intel-servers in hetzelfde cluster te laten draaien en dus twee verschillende CPU's te gebruiken," zegt Sjerps. "Je krijgt alleen performanceproblemen, omdat de langzaamste server de runtime van de query's bepaalt. Maar technisch kan het gewoon. Dat betekent dat de gebruiker langzaam kan groeien en de oudere servers in de loop der tijd kan vervangen."

MapReduce

Greenplum gebruikt ook de technologie die MapReduce genoemd wordt, een technologie die samen met Google werd ontwikkeld. "Google geeft je in een oogwenk antwoord op je vragen. Hoe doen ze dat zo snel? Google weet toch niet wat je volgende vraag zal zijn? Ze gebruiken een technologie met een kleine string: in milliseconden pingt die duizenden servers over de hele wereld. In dat hele kleine moment werkt Google alleen, exclusief voor jou. En toont je de resultaten," legt Weimar uit. "Als je informatie toevoegt aan je query duurt het iets langer, omdat eerst het aantal servers moet worden teruggebracht. Nu gebruikt Google daarnaast een technologie die als het ware parallelle ping mogelijk maakt. Die hele technologie wordt MapReduce genoemd en past heel goed bij onze MPP. MapReduce wordt ook toegepast door andere leveranciers, maar wij zijn gewoon de eerste die het daadwerkelijk geïmplementeerd heeft."

Greenplum kan binnenkort ook draaien op een virtuele machine

Greenplum is géén column-based database. Het bedrijf ontwikkelde in 2009 een inmiddels beschermde technologie die 'polymorphic storage' heet. Weimar: "Dat is superslimme technologie. We slaan parallel op zowel in rijen als in kolommen. Het lijkt een beetje op een aantal Excelsheets op een rij: je kunt de kubus razendsnel in plakjes snijden. Dat is echt geweldig. We haalden er het wereldrecord laden mee en hebben er inmiddels patent op. Deze technologie gaf EMC het laatste zetje tijdens de overnamegesprekken."

Analytics Lab

Nog dit najaar brengt Greenplum een cloudoplossing op de markt, Chorus genaamd. Er zijn al klanten die deze technologie gebruiken. "Het is goed toe te passen als private cloud voor service providers, omdat die hardware op verschillende locaties op de wereld hebben draaien die op enig moment even niets te doen hebben en daar alleen maar staan te staan, met een open webverbinding. Met Chorus kunnen ze elk type data op deze machines laden en ze gebruiken. We zijn daarmee ver vooruit, we zijn er al meer dan twee jaar mee bezig en het product is erg volwassen. En we hebben al klanten die het gebruiken. We zijn de eerste die dit kan aanbieden, het is een echte doorbraak. Maar we hebben meer te bieden. Greenplum komt als software, of als appliance, wat je maar wilt. We kunnen zelfs een gecustomiseerde appliance bouwen – binnen een uur in onze fabriek. Niemand kan dat. En Greenplum kan binnenkort ook draaien op een virtuele machine."

Dan is er nog Greenplum Analytics Lab. "Dat is geen product maar een service, zegt Bart Sjerps. "Ik zal het uitleggen. Er is een groep mensen, sommigen in de VS anderen in Europa. Deze

Uwe Weimar heeft al een 25-jarige carrière in de IT-markt achter de rug, waarvan hij zich de laatste 15 jaar specialiseerde in datawarehousing en Business Intelligence. Hij kent de BI-markt dus erg goed en kan bovendien zijn zegje doen als het over technologie gaat. Weimar was nauw betrokken bij het opzetten van de Cognosvestiging in Duitsland en vertrok daarna naar San Mateo, California, waar hij een van de oprichters was van Active Technologies. Dat bedrijf ontwikkelde een succesvol ETL-tool, en werd daarom overgenomen door BusinessObjects - en is nu dus in de handen van SAP. "Ons toenmalige product heet nu dus Data Integrator," vertelt hij. "Het is goed om te zien dat dat product het vandaag de dag nog steeds zo goed doet."

Een ander bedrijf waar Weimar voor werkte is Hummingbird – nu OpenText - en hij werkte enige jaren voor New Frontiers in Nieuwegein, waarvoor hij de Oost-Europese markt ontgon. "Het is waar, er zijn vele database-bedrijven in de buurt van San Mateo, zoals Aster Data. Het is grappig om te zien hoe ze allemaal technologie van elkaar lenen en vervolgens zelf een nieuw bedrijf beginnen. Ik denk dat dit het Amerikaanse antwoord is op de eeuwenoude Chinese kunst van het kopiëren."

mensen kunnen klanten helpen analytics te doen op hun databases en ze zijn allemaal experts op dit gebied. Niet elk bedrijf heeft mensen in dienst die experts zijn in statistieken, marketing, data mining en *information discovery*. Deze bedrijven zullen zich dus afvragen hoe ze ooit goede analytics kunnen ontwikkelen. Zo'n bedrijf kan contact opnemen met Greenplum Analytics Lab en de mensen daar uitleggen wat de doelstellingen en wensen zijn. Vervolgens worden de data naar het Lab gestuurd, dat een soort POC draait en de resultaten bediscussieert. En bouwt van daar verder. Het is misschien een wat ongewone service van een softwareleverancier, maar wij denken dat het een beetje raar is om een analytische database te hebben en niet te weten wat je er allemaal mee kunt doen."

Weimar voegt daaraan toe: "Als het bedrijf het daar mee eens is, kunnen we het nieuw ontwikkelde analytische model namens hen verkopen aan derden. We hebben al veel van die modellen, maar ze moeten altijd worden gecustomized."

Point-of-sale analytics

Greenplum haakt ook makkelijk aan bij nieuwe technologieën, zoals Hadoop. Greenplum nam Hadoop vanaf het begin serieus. "Hadoop kan omgaan met ongestructureerde data," zegt Weimar. "Bovendien is het een trend in de markt om datawarehousing real-time of toch minstens near real-time te doen. Denk wat de combinatie van Hadoop en Greenplum allemaal kan." Het voorbeeld bij uitstek is een grote klant in Frankrijk, een retailer. American Express heeft een dochtermaatschappij, Loyalty Partners, en samen brengen ze *point-of-sale analytics* op de markt. "Dat is buitengewoon interessant. Stel, je koop een nieuw overhemd. Je betaalt het met je creditcard en de winkelier

vertelt je dat er volgende week een nieuw model broek in de winkel komt, die perfect bij het overhemd past. Op de toonbank staat een soort aanbevelingsdesk, die je producten uit de winkel aanbiedt die je zouden kunnen interesseren. Niet alleen producten die er al zijn, maar ook die nog gaan komen. Ze hadden al een systeem dat dit kon, maar het probleem was dat het er 35 seconden over deed. Precies de tijd waarin je normaal gesproken na het afrekenen de winkel verlaat. Nu gebruiken ze Greenplum/Hadoop, en het promotieresultaat komt nu binnen 7 seconden. En dat is dus gebaseerd op ongestructureerde data, historische data, wat voor type klant je bent, welke kleding je eerder in de winkel gekocht hebt – of online.”

SAS

Op dit moment zijn gesprekken gaande tussen Greenplum en SAS. Het gaat om een lange-termijn strategie, want Greenplum wil niet voortdurend switchen van partner. “SAS is marktleider als het gaat om *deep analytics* en we hopen nieuwe markten te kunnen aanboren samen met SAS.” Uwe Weimar kan het niet laten nog even terug te komen op Teradata. “De mensen hebben echt genoeg van het licentiemodel van Teradata. Je koopt hun appliances helemaal niet, in feite lease of huur je ze. Dus moet je elk jaar een flink bedrag ophoesten. Waarom ik dit zeg?

Onze grootste doelmarkt is het vervangen van Teradata door Greenplum. En daar hebben we SAS voor nodig, zij hebben de kennis en de skills. En wij hebben de bijbehorende technologie.”

Bart Sjerps kijkt nog even naar de Nederlandse markt. “We beginnen net de markt kennis te laten maken met Greenplum. De afgelopen drie jaar was ik bij EMC de Oracle-expert. Toen Oracle met Exadata kwam, beschouwde EMC dat als een concurrerend product. Het was logisch, dat toen EMC Greenplum overnam, men aan mij vroeg me daar eens in te verdiepen. Ik ben direct begonnen met mijn EMC-collega's te kijken naar de mogelijkheden van cross-selling. We hebben inmiddels al klanten in Nederland en we zijn nu met onze partners in overleg hoe we dat aantal kunnen laten groeien. Nogmaals: we beginnen net.” “Nederland is een belangrijke markt voor ons. We zoeken op dit moment ook naar goede mensen,” besluit Weimar. “Ik manage Greenplum nu in 22 landen, in Centraal- en Oost-Europa. We willen 20 procent van de totale omzet in Nederland gaan halen. Waarom? Omdat hier goede markten zijn voor retail, olie en gas en telco's. Sommigen noemen de Nederlandse markt conservatief. Nou, de Duitse markt is veel conservatiever, dat verzeker ik je.”

Hans Lamboo is freelance journalist.

Database Magazine vacatures op Array Jobs

Hieronder staat een selectie van vacatures die tevens op de site van Database Magazine te vinden zijn en passen bij jouw functieprofiel. Het volledige vacatureaanbod is op www.arrayjobs.nl te vinden. Array Jobs is een nieuwe dienst van Array Publications die deze zomer live is gegaan. Maak zo snel mogelijk een eigen account aan en sla je zoekopdrachten op.

Benut de mogelijkheden van onze Job Search Agent en vacatures per rss feed. Niet alleen zoeken maar ook snel gevonden worden! Dit vergroot de kans op een uitdagende nieuwe baan. **Plaats jouw CV op Array Jobs!**

Solution Designer (interim)

Locatie: Amsterdam. Werkgever: Independent Recruiters

Datawarehouse Lead Consultant

Locatie: Amsterdam. Werkgever: Independent Recruiters

Diverse SAS Consultants

Locatie: Noord-Holland/Gelderland. Werkgever: Jenrick

Business Intelligence Informatie analist

Locatie: Rijswijk. Werkgever: TopBl

Datawarehouse specialist

Locatie: Rijswijk. Werkgever: TopBl

Lead Technisch Applicatie Beheerder

Locatie: Amsterdam. Werkgever: Independent Recruiters

Oracle DBA (Senior)

Locatie: Almere. Werkgever: Independent Recruiters

Datamanagement specialist / BI consultant

Locatie: Randstad. Werkgever: Recruitment 4 you Almere

Oracle Technisch Applicatie Beheerder (Lead)

Locatie: Amsterdam. Werkgever: Independent Recruiters

Business Objects Reporting Specialist

Locatie: Utrecht. Werkgever: Independent Recruiters

Business Objects Specialist

Locatie: Amsterdam. Werkgever: Independent Recruiters

Senior System Engineer

Locatie: Huizen. Werkgever: heeftictvacatures.nl

SAP Finance Banking and Insurance Consultant

Locatie: Utrecht. Werkgever: Independent Recruiters

Application Engineer Business Objects

Locatie: Amsterdam. Werkgever: Independent Recruiters

SQL specialist

Locatie: Amsterdam. Werkgever: Klikwork

System Administrator

Locatie: Den Haag. Werkgever: Independent Recruiters

Business Intelligence Consultant

Locatie: Amersfoort. Werkgever: Novaccent

DBA Consultant

Locatie: Amersfoort. Werkgever: Novaccent

Oracle DBA

Locatie: Amersfoort. Werkgever: Novaccent

Kijk op www.arrayjobs.nl voor meer vacatures!